Bridging logistic and OLS regression

Constantine Kapsalis

Data Probe Economic Consulting

April 2010

DATA PROBE ECONOMIC CONSULTING INC.

9 Maki Place, Ottawa, Canada K2H 9R5
www.dataprobeinc.ca dataprobeinc@hotmail.com

BRIDGING LOGISTIC AND OLS REGRESSION

Constantine Kapsalis, Ph.D.

Working Papers Series: No. 2011-1

January 1, 2011

Abstract

There is broad consensus that logistic regression is superior to ordinary least squares (OLS) regression at predicting the probability of an event. OLS is still widely used in binary choice models because its coefficients are easier to interpret, while the resulting estimates tend to be close to the logit estimates anyway. Although some statistical software provide an easy way of calculating marginal effects (equivalent in interpretation to OLS coefficients) this is not always the case. This paper shows a simple way of calculating marginal effects from logistic coefficients.

The author would appreciate receiving comments. Please email your comments to: kapsalis@sympatico.ca

1. Introduction

There are several instances in economic studies where the dependent variable is not continuous but dichotomous (e.g. labour force participation, unemployment, poverty, reliance on social assistance). In these situations, the more familiar OLS regression has limitations and a logistic regression, or its very similar probit regression, is a better choice. Specifically, the two main limitations of OLS are: (a) fitted values of the dependent variable (*P)* could fall outside the zero-one range; and (b) the error term *e* is heteroskedastic (Goldberger, 1964; Theil, 1981).

Unfortunately, logistic regression coefficients do not have the same intuitive interpretation as OLS coefficients do. In the case of OLS the dependent variable is the probability of the event itself:

(1) $$p = b_0 + \sum b_i X_i + u$$

where *p* is the binary dependent variable (taking the values of 1 or zero), b_i*'s* are the OLS linear coefficients, X_i*'s* are the independent variables, and *u* is the error term. For example, if the binary event is being unemployed or not and X_i refers to the female gender, then the b_i coefficient shows how much more likely are females to experience unemployment than males, keeping all other attributes the same.

By contrast, in the case of logistic regression the dependent variable is not the probability of the event but its logistic transformation:

(2) $$\ln\left(\frac{p}{1-p}\right) = b_0 + \sum b_i X_i$$

Consequently, the b_i coefficients show the impact of each independent variable not on the probability of the event itself, but on its logistic transformation. The problem now is that, although the logistic model is more appropriate than OLS, we are left with regression coefficients that are difficult to interpret intuitively.

As a result, many practitioners recommend the OLS model as an approximation of the more correct logistic model or as a preliminary analysis tool (Moffit, 1999; Amemiya, 1981). This approach has been reinforced by the fact that the two models tend to lead to similar results, at least in terms of the partial derivatives of the dependent probability with respect to individual independent variables (Pohlmann and Leitner, 2003).

2. Estimating Marginal Effects

An alternative approach to relying on OLS is to derive the from the logistic regression results the marginal effects of changes in the independent variables on the dependent variable. This way we can combine the superior statistical properties of logistic regression with the intuitive nature of OLS coefficients.

Some software, like STATA, provide as an option the calculation of marginal effects. Unfortunately this is not the case with all statistical software (like SPSS). One approach that has been used in the literature is to estimate marginal effects by comparing point estimates of the expected probability of various characteristics (Pohlmann and Leitner, 2003). For example, the marginal effect of the impact of female gender on the probability of unemployment can be derived from the results of a logistic regression by estimating the female and male probabilities, keeping the values of the rest of independent variables equal to their average value, and subtracting the two. Of course, since the relationship is non-linear, the results will tend to differ depending on the choice of the point where the partial derivatives are estimated and the degree of non-linearity of the relationship.

The difficulty with the above approach is that it is computationally demanding. However, there is a simpler way of estimating marginal effects from a logistic regression using the odds ratio.

The odds ratio is simply the exponential value of the logit coefficients. In logistic regression, odds are defined as the ratio $p/(1-p)$ and the odds ratio (R) is defined as the ratio of two odds:

$$R = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \tag{3}$$

where p_1 refers to the probability of an event (e.g. unemployment) for a particular characteristic (e.g. females) and p_0 refers to the corresponding probability of the omitted characteristic (in this case male). By solving the above equation for p_1 and assigning a specific value to p_0 we can easily estimate the corresponding marginal effect (M):

$$M = p_1 - p_0 = \left(\frac{Rp_0}{1-p_0+Rp_0} \right) - p_0 \tag{4}$$

In the case of dummy independent variables, p_0 will be the average probability of the omitted category. Using the previous example, in the case of gender the marginal effect will show the impact of being female on the probability of unemployment, keeping the rest of the rest of the female characteristics the same as those of males. In the case of a continuous independent variable (e.g. age) p_0 can be simply set equal to the overall average unemployment rate of the data sample.

3. An Example

We now present a simple example to illustrate the proposed methodology. The dependent variable is the probability of experiencing unemployment during the year among those who were in the labour force for at least part of the year. The independent variables include a continuous one (age) and several dummy variables (gender, education, province, area, and disability). The source of data is Statistics Canada's Survey of Labour and Income Dynamics (SLID), 2007. The sample includes 35,061 labour force participants, age 18-64.

Table 1 presents the standard SPSS regression results for OLS and logistic regression. The last column shows the marginal effects, based on the logistic regression results. In addition to illustrating the method of estimating marginal effects, Table 1 reconfirms the finding in the literature that logistic and OLS regression results tend to be similar. In the case of the particular example, virtually all OLS coefficients were within one percentage point of the corresponding marginal effects that were based on the logistic regression results.

4. Conclusion

This paper has presented a simple way of estimating marginal effects from logistic regression results. It has also demonstrated with an example that OLS coefficients tend to be very close to logistic marginal effects. Thus the paper provides analysts a simple way of combining the benefits of using logistic regression with the practical advantage of producing intuitive coefficients that are easier to communicate to a broader audience. The proposed method simply requires the calculation of the average value of the dependent variable for each of the omitted categories and, using these values along with the odds ratio (which are simply the exponentials of the logit b coefficients), solve equation (4).

Table 1. OLS versus logit regression estimates of the rate of unemployment

Independent variables	OLS regression			Logit regression				
	b-coef.	Std err.	t-stat.	b-coef.	Std err.	t-stat.	Odds ratio	Marginal effect
Constant	0.520	0.021	24.382	0.877	0.148	5.938	2.403	
Age (continuous)	-0.005	0.000	-32.021	-0.040	0.001	-30.511	0.961	-0.005
Sex								
- Male (omitted)								
- Female	0.000	0.004	0.077	0.007	0.030	0.244	1.007	0.001
Education								
- Less than 9 years (omitted)								
- 9-10 years	0.020	0.015	1.366	0.110	0.102	1.083	1.117	0.018
- 11-13 years	0.041	0.015	2.759	0.091	0.102	0.898	1.096	0.015
- High school diploma	-0.054	0.013	-4.258	-0.401	0.092	-4.377	0.670	-0.056
- Some college	-0.029	0.013	-2.190	-0.270	0.094	-2.878	0.763	-0.040
- Some university	-0.019	0.014	-1.373	-0.219	0.099	-2.224	0.803	-0.033
- College certificate	-0.077	0.012	-6.440	-0.576	0.087	-6.653	0.562	-0.077
- University BA	-0.111	0.013	-8.750	-0.881	0.095	-9.276	0.414	-0.106
- University above BA	-0.109	0.014	-7.845	-0.916	0.110	-8.312	0.400	-0.109
Province								
- Newfoundland (omitted)								
- PEI	-0.011	0.033	-0.320	-0.071	0.220	-0.322	0.932	-0.013
- Nova Scotia	-0.049	0.020	-2.495	-0.287	0.134	-2.152	0.750	-0.052
- New Brunswick	-0.048	0.021	-2.312	-0.267	0.141	-1.895	0.766	-0.048
- Quebec	-0.076	0.017	-4.606	-0.464	0.111	-4.184	0.629	-0.079
- Ontario	-0.094	0.016	-5.686	-0.597	0.110	-5.415	0.551	-0.098
- Manitoba	-0.140	0.019	-7.345	-0.985	0.138	-7.122	0.373	-0.145
- Saskatchewan	-0.116	0.020	-5.906	-0.789	0.141	-5.610	0.454	-0.123
- Alberta	-0.134	0.017	-7.836	-0.932	0.117	-7.959	0.394	-0.139
- BC	-0.108	0.017	-6.394	-0.722	0.115	-6.277	0.486	-0.115
Area								
- Rural (omitted)								
- Urban: 0 to 29,999	-0.006	0.009	-0.733	-0.036	0.064	-0.564	0.965	-0.005
- Urban: 30,000 to 99,999	-0.014	0.009	-1.521	-0.096	0.069	-1.395	0.908	-0.014
- Urban: 100,000 to 499,999	-0.024	0.008	-2.895	-0.171	0.061	-2.793	0.843	-0.025
- Urban: 500,000 and higher	-0.007	0.007	-0.955	-0.046	0.054	-0.846	0.955	-0.007
Disability								
- No (omitted)								
- Yes	0.073	0.005	14.481	0.535	0.037	14.551	1.708	0.083
OLS R^2/ Logit Nagelkerke R^2	0.054			0.091				
Number of cases	35,061			35,061				

References

Amemiya, T. (1981). "Qualitative response models: a survey", *Journal of Economic Literature* 19: 1483-1536.

Goldberger, A. (1964). *Econometric theory* (Wiley, New York).

Moffitt, Robert (1999). "New Developments in Econometric Methods for Labor Market Analysis", in *Handbook of Labour Economics*, Volume 3, Chapter 24, Edited by O. Ashenfelter and D. Card.

Pohlmann, John T. and Dennis W. Leitner (2003). "A Comparison of Ordinary Least Squares and Logistic Regression", *The Ohio Journal of Science*, 103 (5): 118-125.

Theil, H. (1981). *Principles of econometrics* (Wiley, New York).